A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series

Demetris Koutsoyiannis

Department of Water Resources, Faculty of Civil Engineering, National Technical University, Athens

Abstract. A generalized framework for single-variate and multivariate simulation and forecasting problems in stochastic hydrology is proposed. It is appropriate for short-term or long-term memory processes and preserves the Hurst coefficient even in multivariate processes with a different Hurst coefficient in each location. Simultaneously, it explicitly preserves the coefficients of skewness of the processes. The proposed framework incorporates short-memory (autoregressive moving average) and long-memory (fractional Gaussian noise) models, considering them as special instances of a parametrically defined generalized autocovariance function, more comprehensive than those used in these classes of models. The generalized autocovariance function is then implemented in a generalized moving average generating scheme that yields a new time-symmetric (backward-forward) representation, whose advantages are studied. Fast algorithms for computation of internal parameters of the generating scheme are developed, appropriate for problems including even thousands of such parameters. The proposed generating scheme is also adapted through a generalized methodology to perform in forecast mode, in addition to simulation mode. Finally, a specific form of the model for problems where the autocorrelation function can be defined only for a certain finite number of lags is also studied. Several illustrations are included to clarify the features and the performance of the components of the proposed framework.

1. Introduction

Since its initial steps in the 1950s, stochastic hydrology, the application of theory of stochastic processes in analysis and modeling of hydrologic processes, has offered very efficient tools in tackling a variety of water resources problems, including hydrologic design, hydrologic systems identification and modeling, hydrologic forecasting, and water resources management. An overview of the history of stochastic hydrology has been compiled by *Grygier and Stedinger* [1990]. We mention as the most significant initial steps of stochastic hydrology the works by *Barnes* [1954] (generation of uncorrelated annual flows at a site from normal distribution), *Maass et al*. [1962] and *Thomas and Fiering* [1962] (generation of flows correlated in time), and *Beard* [1965] and *Matalas* [1967] (generation of concurrent flows at several sites). We must mention that the foundation of stochastic hydrology followed the significant developments in mathematics and physics in the 1940s, as well as the development of computers. Specifically, it followed the establishment of the Monte Carlo method, which was invented by Stanislaw Ulam in 1946 (initially conceived as a method to estimate probabilities of solitaire success; soon after, it grew to solve neutron diffusion problems), developed by himself and other great mathematicians and physicists in Los Alamos (John von Neumann, Nicholas Metropolis, Enrico Fermi), and first implemented on the electronic numerical integrator and computer (ENIAC) [*Metropolis*, 1989; *Eckhardt*, 1989].

The classic book on time series analysis by *Box and Jenkins* [1970] also originated from different, more fundamental scientific fields. However, it has subsequently become very important in stochastic hydrology and still remains the foundation of hydrologic stochastic modeling. Box and Jenkins developed a classification scheme for a large family of time series models. Their classification scheme distinguishes among autoregressive models of order p (AR(p)), moving average models of order q (MA(q)), and combinations of the two, called autoregressive moving average (ARMA(p, q)).

However, despite making a large family, all Box-Jenkins models are essentially of short-memory type; that is, their autocorrelation structure decreases rapidly with the lag time. Therefore such models are proven inadequate in stochastic hydrology, if the long-term persistence of hydrologic (and other geophysical) processes is to be modeled. This property, discovered by *Hurst* [1951], is related to the observed tendency of annual average streamflows to stay above or below their mean value for long periods. Other classes of models such as fractional Gaussian noise (FGN) models [*Mandelbrot*, 1965; *Mandelbrot and Wallis*, 1969a, b, c], fast fractional Gaussian noise models [*Mandelbrot*, 1971], and broken line models [*Ditlevsen*, 1971; *Mejia et al*., 1972] are more appropriate to resemble long-term persistence [see also *Bras and Rodriguez-Iturbe*, 1985, pp. 210–280]. However, models of this category have several weak points such as parameter estimation problems, narrow type of autocorrelation functions that they can preserve, and their inability to perform in multivariate problems (apart from the broken line model, see *Bras and Rodriguez-Iturbe* [1985, p. 236]). Therefore they have not been implemented in widespread stochastic hydrology packages such as LAST [*Lane and Frevert*, 1990], SPIGOT [*Grygier and Stedinger*, 1990], CSUPAC1 [*Salas*, 1993], and WASIM [*McLeod and Hipel*, 1978].

Another peculiarity of hydrologic processes is the skewed distribution functions observed in most cases. This is not so

Paper number 2000WR900044.
0043-1397/00/2000WR900044$09.00

common in other scientific fields whose processes are typically Gaussian. Therefore attempts have been made to adapt standard models to enable treatment of skewness [e.g., *Matalas and Wallis*, 1976; *Todini*, 1980; *Koutsoyiannis*, 1999a, b]. (The presentation by *Koutsoyiannis* [1999b] is available on-line from World Wide Web server for the National Technical University, Athens, node at http://www.hydro.ntua.gr/faculty/dk/pub/pskewness.pdf.) The skewness is mainly caused by the fact that hydrologic variables are nonnegative and sometimes have an atom at zero in their probability distributions. Therefore a successful modeling of skewness indirectly contributes to avoiding negative values of simulated variables; however, it does not eliminate the problem, and some ad hoc techniques (such as truncation of negative values) are often used in addition to modeling skewness.

The variety of available models, either short memory or long memory, with different equations, parameter estimation techniques, and implementation, creates difficulties in the model choice and use. Besides, the AR(p) or ARMA(p, q) models, which have been more widespread in stochastic hydrology, become more and more complicated, and the parameters to be estimated become more uncertain, as p or q increases (especially in multivariate problems). Thus in software packages such as those mentioned above, only AR(0) through AR(2) and ARMA(1, 1) models are available.

The reason for introducing several models and classifying them into different categories seems to be not structural but rather imposed by computational needs at the time when they were first developed. Today, the widespread use of fast personal computers allows a different approach to stochastic models. In this paper, we try to unify all the above-described models, both short memory and long memory, simultaneously modeling the process skewness explicitly. The unification is done using a generalized autocovariance function (section 2), which depends on a number of parameters, not necessarily greater than that typically used in traditional stochastic hydrology models. Specifically, we separate the autocovariance function from the mathematical structure of the generating scheme (or model) that implements this autocovariance function. Thus the autocovariance function may depend on two or three parameters, but the generating scheme may include a thousand numerical coefficients (referred to as internal parameters), all dependent on (and derived from) these two or three parameters. The generating scheme used is of moving average type, which is the simplest and most convenient; in addition to the traditional backward moving average scheme, a new scheme with several advantages, referred to as symmetric (backward-forward) moving average model, is introduced (section 3). New methods of estimating the internal parameters of the generating scheme, given the external parameters of the autocorrelation function, are introduced (section 4); they are very fast even for problems including thousands of internal parameters. The proposed generating scheme can be directly applied for stochastic simulation. In addition, the scheme can perform in forecast mode, as well, through a proposed methodology that makes this possible (section 5); it is a generalized methodology that can be used with any type of stochastic model. The framework, initially formulated as a single-variate model, is directly extended for multivariate problems (section 6). A specific model form for problems where the autocorrelation function is defined only for a certain finite number of lags (e.g., in generation of rainfall increments within a rainfall event) is also studied (section 7). In its present form the proposed framework is formulated for stationary processes; the possibility of incorporating seasonality in combination with seasonal models is also mentioned briefly (section 8, also including conclusions). Several sections of the paper include simple illustrations that clarify the features and the performance of the components of the proposed framework. Additional examples on the application of the generalized autocovariance function using synthetic and historical hydrologic data sets are given in Appendix A1.[1] To increase readability, several mathematical derivations are given separately in Appendices A2–A4.

2. A Generalized Autocovariance Structure and Its Spectral Properties

Annual quantities related to hydrologic processes such as rainfall, runoff, evaporation, etc. or submonthly quantities of the same processes (e.g., fine-scale rainfall depths within a storm) can be modeled as stationary stochastic processes in discrete time. We consider the stationary stochastic process X_i in discrete time denoted with i, with autocovariance

$$\gamma_j := \mathrm{Cov}\,[X_i, X_{i+j}] \qquad j = 0, 1, 2, \ldots \tag{1}$$

The variables X_i are not necessarily standardized to have zero mean or unit variance, nor are they necessarily Gaussian; on the contrary, they can be skewed with coefficient of skewness $\xi_X := E[(X_i - \mu_X)^3]/\gamma_0^{3/2}$, where $\mu_X := E[X_i]$ is the mean and γ_0 is the variance. The skewness term, which is usually ignored in stochastic process theory, is essential for stochastic hydrology because hydrologic variables very often have skewed distributions. The parameters μ_X, γ_0, and ξ_X determine in an acceptable approximation the marginal distribution function of the hydrologic variable of interest, whereas the autocovariances γ_j, if known, determine, again in an acceptable approximation, the stochastic structure of the process. However, as γ_j is estimated from samples $x_1, \ldots, x_n$ with typically small length n, only few foremost terms γ_j can be known with some acceptable confidence. Usually, these are determined by the (biased) estimator [e.g., *Bloomfield*, 1976, pp. 163, 182; *Box and Jenkins*, 1970, p. 32; *Salas*, 1993, p. 19.10]

$$\hat{\gamma}_j = \frac{1}{n}\sum_{i=1}^{n-j}(x_i - \bar{x})(x_{i+j} - \bar{x}), \tag{2}$$

where $\bar{x}$ is the sample mean. In addition to the fact that the number of available terms of the sum in (2) decreases linearly with the lag j (which results in increasing estimation uncertainty), typically, γ_j is a decreasing function of lag j. The combination of these two facts may lead us to consider that γ_j is zero beyond a certain lag m (i.e., for $j \geq m$) which may be not true. In other words, the process X_i may be regarded as short memory, while in reality it could be long memory. However, the large lag autocovariance terms γ_j may affect seriously some properties of the process of interest, and thus a choice of a short-memory model would be an error as far as these properties are considered. This is the case, for example, if the properties of interest are the duration of droughts or the range of cumulative departures from mean values [e.g., *Bras and Rodriguez-Iturbe*, 1985, pp. 210–211].

[1] Supporting Appendices A1–A4 are available with entire article on microfiche. Order by mail from American Geophysical Union, 2000 Florida Avenue, N. W., Washington, DC 20009 or by phone at 800-966-2481; $2.50. Document 2000WR900044M. Payment must accompany order.

The most typical stochastic models, belonging to the class of ARMA(p, q) models [*Box and Jenkins*, 1970] can be regarded as short-memory models (although the ARMA(1, 1) model has been used to approximate long-term persistence for some special values of its parameters [*O'Connell*, 1974]) as they essentially imply an exponential decay of autocovariance. Specifically, in an ARMA process, the autocovariance for large lags j converges either to

$$\gamma_j = a\rho^j \tag{3}$$

if all terms γ_j are positive or to

$$\gamma_j = a\rho^j \cos(\theta_0 + \theta_1 j) \tag{4}$$

if the terms alternate in sign, where a, ρ, θ_0, and θ_1 are numerical constants (with $0 \le \rho \le 1$). The case implied by (3) is more common than (4) if the process X_i represents some hydrologic quantity like rainfall, runoff, etc.

The inability of the ARMA processes to represent important properties of hydrologic processes, such as those already mentioned, have led *Mandelbrot* [1965] to introduce another process known as fractional Gaussian noise (FGN) process [see also *Bras and Rodriguez-Iturbe*, 1985, p. 217]. This is a long-memory process with autocovariance

$$\gamma_j = \gamma_0\{(1/2)[(j-1)^{2H} + (j+1)^{2H}] - j^{2H}\} \tag{5}$$
$$j = 1, 2, \dots,$$

where H is the so-called Hurst coefficient $(0.5 \le H \le 1)$. Apart from the first few terms, this function is very well approximated by

$$\gamma_j = \gamma_0(1 - 1/\beta)(1 - 1/2\beta)j^{-1/\beta}, \tag{6}$$

where $\beta = 1/[2(1-H)] \ge 1$, which shows that autocovariance is a power function of lag. Notably, (5) is obtained from a continuous time process $\Xi(t)$ with autocovariance Cov $[\Xi(t), \Xi(t+\tau)] = a\tau^{-1/\beta}$ (with constant $a = \gamma_0(1 - 1/\beta)(1 - 1/2\beta)$), by discretizing the process using time intervals of unit length and taking as X_i the average of $\Xi(t)$ in the interval $[i, i+1]$.

The autocovariances of both ARMA and FGN processes for large lags can be viewed as special cases of a generalized autocovariance structure (GAS)

$$\gamma_j = \gamma_0(1 + \kappa\beta j)^{-1/\beta}, \tag{7}$$

where κ and β are constants. Indeed, for $\beta = 0$, (7) becomes (using de l'Hospital's rule)

$$\gamma_j = \gamma_0 \exp(-\kappa j), \tag{8}$$

which is identical to (3) if $\kappa = -\ln\rho$. For $\beta > 1$ and large j, (7) yields a very close approximation of (6) if

$$\kappa = 1/\{\beta[(1 - 1/\beta)(1 - 1/2\beta)]^\beta\} =: \kappa_0. \tag{9}$$

For other values of κ or for values of β in the interval (0, 1), (7) offers a wide range of feasible autocovariance structures in between, or even outside of, the ARMA and the FGN structures, as demonstrated in Figure 1a, where we have plotted several autocovariance functions using different values of β but keeping the same γ_0, γ_1, and γ_2 for all cases. The meaning of the different values of β will be discussed later, in the end of this section. Here it may suffice to explain that GAS is more comprehensive than the FGN scheme as the latter, with its single parameter H, cannot model explicitly even the lag-one autocovariance. It is also more comprehensive than ARMA schemes as it can explicitly model long-term persistence while yet being parameter parsimonious.

If the autocovariance is not everywhere positive but alternates in sign, (7) can be altered in agreement with (4) to become

$$\gamma_j = \gamma_0(1 + \kappa\beta j)^{-1/\beta} \cos(\theta_0 + \theta_1 j). \tag{10}$$

In the form of (7) or (10), GAS has three or five parameters, respectively, one of which is the process variance γ_0 (thus the corresponding autocorrelation structure has two or four parameters, respectively). Although parameter parsimony is most frequently desired in stochastic modeling [e.g., *Box and Jenkins*, 1970, p. 17], GAS can be directly extended to include a greater number of parameters. Specifically, it can be assumed that the initial $m + 1$ terms γ_j $(j = 0, \dots, m)$ have any arbitrary values (e.g., estimated from available records), and then (7) (or (10)) is used for extrapolating for large j. Essentially, this introduces m additional model parameters at most. Thus the total number of independent parameters is $m + 1$ if both κ and β are estimated in terms of $\gamma_0, \dots, \gamma_m$, or $m + 2$ if β is estimated independently; even in the latter case, κ cannot be regarded as an independent parameter because continuity at term γ_m demands that

$$\kappa = \begin{cases} \dfrac{1}{\beta m}\left[\left(\dfrac{\gamma_0}{\gamma_m}\right)^\beta - 1\right] & \beta > 0, \\ \dfrac{1}{m}\ln\left(\dfrac{\gamma_0}{\gamma_m}\right) & \beta = 0. \end{cases} \tag{11}$$

Parameter estimation can be based on the empirical autocovariance function. In the most parameter parsimonious model form (7), parameter γ_0 is estimated from the sample variance, and parameters κ and β can be estimated by fitting GAS to the empirical autocovariance function. There are several possibilities to fit these parameters: (1) We can choose to have a good overall fit to a number of autocovariances without preserving exactly any specific value. (2) Alternatively, we may choose to preserve exactly the lag-one autocovariance γ_1 in which case (11) holds for $m = 1$. Still we have an extra degree of freedom (one more independent parameter), which can be estimated so as to get a good fit of GAS to autocovariances for a certain number of lags higher than 1. (3) Finally, we may choose to preserve the lag-one and lag-two autocovariances exactly. Case 3 is the easiest to apply, as parameters κ and β are directly estimated from (7) in terms of γ_0, γ_1, and γ_2. However, cases 1 and 2 are preferable because they take into account the autocovariances of higher lags and, thus, the long-memory properties of the process. A least squares method is a direct and simple basis to take into account more than two autocovariances in cases 1 and 2. Note that linearization of (7) by taking logarithms is not applicable as some empirical autocovariances may be negative. Therefore the application of least squares requires nonlinear optimization, which is a rather simple task as there are only two parameters in case 1 (κ and β) and one in case 2 (only β because of (11)). Results of applications of method 1 on synthetic and historical hydrologic data sets are given in Appendix A1. Apparently, parameter estimation is subject to high uncertainty, and this is particularly true for β, which is related to the long-term persistence of the process. Therefore, if the record length is small, it may be a good idea to assume a value of β after examining other series

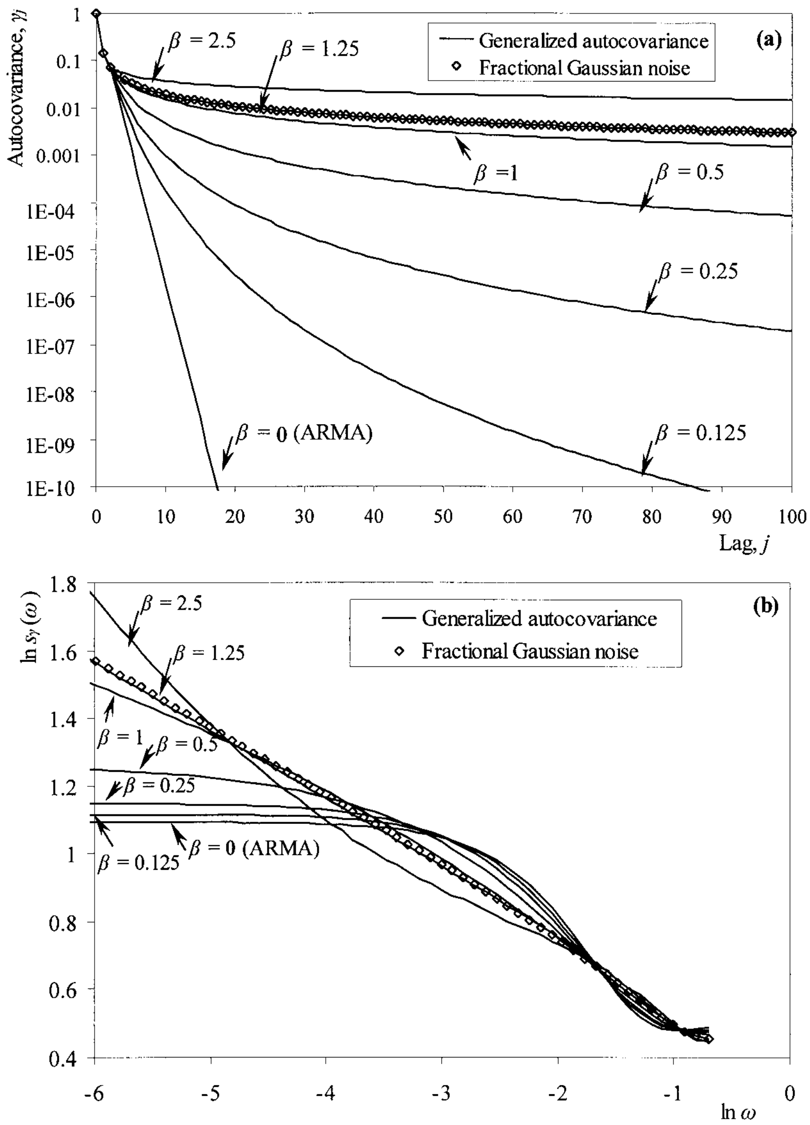

Figure 1. (a) Examples of autocovariance sequences of the proposed generalized type for several values of the exponent β, in comparison with the fractional Gaussian noise and ARMA types; (b) corresponding power spectra. (Read 1E-04 as 1×10^{-4}.)

of nearby gauges, rather than to estimate it directly from the available small record.

Similar parameter estimation strategies can be applied for the richer in parameters forms of GAS that were described above, in which case several values γ_j are preserved. We must emphasize that not any arbitrary sequence γ_j can be a feasible autocovariance sequence. Specifically, γ_j is a feasible autocovariance sequence if it is positive definite [*Papoulis*, 1991, pp. 293–294]. This can be tested in terms of the variance-covariance matrix **h** of the vector of variables $[X_1, \ldots, X_s]^T$, which has size $s \times s$ and entries

$$h_{ij} = \gamma_{|i-j|}. \tag{12}$$

If **h** is positive definite for any s, then γ_j is a feasible autocovariance sequence. Normally, if the model autocovariances are estimated from data records, positive definiteness is satisfied. However, it is not uncommon to meet a case that does not satisfy this condition. The main reason is the fact that autocovariances of different lags are estimated using records of different lengths either due to the estimation algorithm (e.g., using (2)) or due to missing data. Another reason is the fact that high-lag autocovariances are very poorly estimated, as explained above.

An alternative way to test that γ_j is a feasible autocovariance sequence is provided by the power spectrum of the process, which should be positive everywhere. The power spectrum of the process is the discrete Fourier transform (DFT; also termed the inverse finite Fourier transform) of the autocovariance sequence γ_j [e.g., *Papoulis*, 1991, pp. 118, 333; *Bloomfield*, 1976, pp. 46–49; *Debnath*, 1995, pp. 265–266; *Spiegel*, 1965, p. 175]; that is,

$$s_\gamma(\omega) := 2\gamma_0 + 4\sum_{j=1}^{\infty} \gamma_j \cos(2\pi j\omega) = 2\sum_{j=-\infty}^{\infty} \gamma_j \cos(2\pi j\omega). \tag{13}$$

Because γ_j is an even function of j (i.e., $\gamma_j = \gamma_{-j}$), the DFT in (13) is a cosine transform; as usual we have assumed in (13) that the frequency ω ranges in [0, 1/2], so that γ_j is determined in terms of $s_\gamma(\omega)$ by

$$\gamma_j = \int_0^{1/2} s_\gamma(\omega) \cos(2\pi j\omega)\, d\omega. \qquad (14)$$

If autocovariance is given by the generalized relation (7) for all j, then it is easily shown that the power spectrum is

$$s_\gamma(\omega) = 2\gamma_0 \left\{ -1 + 2\left(\frac{1}{\beta\kappa}\right)^{1/\beta} \mathrm{Re}\left[\Phi\left(e^{2i\pi\omega}, \frac{1}{\beta}, \frac{1}{\beta\kappa}\right)\right] \right\} \qquad (15)$$

where $i := \sqrt{-1}$, Re [ ] denotes the real part of a complex number, and $\Phi(\ )$ is the Lerch transcendent phi function defined by

$$\Phi(z, b, a) := \sum_{j=0}^{\infty} \frac{z^j}{(a+j)^b}. \qquad (16)$$

In the specific case that $\beta = 0$, where (8) holds, (15) reduces to

$$s_\gamma(\omega) = 2\gamma_0 \frac{(1 - e^{-2\kappa})}{1 + e^{-2\kappa} - 2e^{-\kappa}\cos(2\pi\omega)}. \qquad (17)$$

This gives a characteristic inverse S-shaped power spectrum (Figure 1b) that corresponds to a short-memory process.

Numerical investigation shows that for any $\beta > 1$ and $\kappa = \kappa_0$ (given in (9)), (15) becomes approximately a power function of the frequency ω with exponent approaching $1/\beta - 1$. (More accurately, the exponent is by an amount δ smaller than $1/\beta - 1$, where δ is ~0.03 for $\beta \geq 2.5$ and decreases almost linearly to zero as β approaches 1. The exponent becomes almost equal to $1/\beta - 1$ if κ is set equal to $[1 + 0.71(1 - 1/\beta)(1 - 1/2\beta)]\kappa_0$; however, in the latter case the departure of the power spectrum from the power law is greater.) This case indicates a typical long-memory process, similar to a FGN process (see Figure 1b) where the power function appears as a straight line on the log-log plot). Generally, the power spectrum tends to infinity as ω tends to zero, regardless of the value of κ, if $\beta \geq 1$. For $\beta < 1$ the power spectrum cannot be a power law of the frequency but approaches that given by (17) (inverse S-shaped) as β decreases, taking a finite value for $\omega = 0$.

If we fix γ_0 and γ_1 (the variance and the lag-one autocovariance) at some certain values, and vary β (and κ accordingly), we observe that there exists a combination of $\beta = \beta^* \geq 1$ and $\kappa = \kappa_0(\beta^*)$ (given by (9)) resulting in a power spectrum $s^*(\omega)$ approximately following a power law. For $\beta > \beta^*$ the power spectrum exceeds $s^*(\omega)$ for low frequencies (that is, it departs from the straight line and becomes inverse J-shaped in the log-log plot). The opposite happens if $\beta < \beta^*$ (the spectrum tends to the inverse S-shaped). This is demonstrated in Figure 1b, where, in addition, γ_2 has been also fixed; the power spectra of Figure 1b are those resulting from the autocovariances of Figure 1a. Note that the power spectra of Figure 1b have been calculated numerically from (13) rather than from (15) because the three fixed autocovariances γ_0, γ_1, and γ_2 do not allow a single instance of (7) to hold for all j.

The case $\beta = \beta^*$ (straight line on log-log plot) that corresponds to the FGN process has been met in many hydrological and geophysical series. The case $\beta = 0$ (inverse S-shaped line on log-log plot) that corresponds to ARMA processes has been widely used in stochastic hydrology. In addition to these cases the GAS scheme allows for all intermediate values of β in the range (0, β^*), as well as for values $\beta > \beta^*$ (inverse J-shaped line on log-log plot, or very "fat" tail of autocovariance). The case $0 < \beta < \beta^*$ implies a long-term persistence weaker than the typical FGN one. The case $\beta > \beta^*$ characterizes processes with strong long-term persistence but not very strong lag-one correlation coefficient. Both these cases can be met in hydrologic series (see examples in Appendix A1).

In section 4.1 we will see how we can utilize the power spectrum of the process to determine the parameters of a generalized generating scheme, which will be introduced in the following section 3.

3. Description of the Generating Scheme

It is well known [*Box and Jenkins*, 1970, p. 46] that for any autocovariance sequence γ_j, X_i can be written as the weighted sum of an infinite number of independent and identically distributed (i.i.d.) innovations V_i (also termed auxiliary or noise variables), that is, in the following form, known as (backward) moving average (BMA) form (where we have slightly modified the original notation of *Box and Jenkins* [1970]):

$$X_i = \sum_{j=-\infty}^{0} a_{-j} V_{i+j} = \cdots + a_2 V_{i-2} + a_1 V_{i-1} + a_0 V_i, \qquad (18)$$

where a_j are numerical coefficients that can be determined from the sequence of γ_j. Specifically, coefficients a_j are related to γ_j through the equation [*Box and Jenkins*, 1970, pp. 48, 81]

$$\sum_{j=0}^{\infty} a_j a_{i+j} = \gamma_i \qquad i = 0, 1, 2, \ldots \qquad (19)$$

Although in theory X_i is expressed in terms of an infinite number of innovations, in practice it suffices to use a finite number of them for two reasons: (1) because the number of variables to be generated in any simulation problem is always a finite number and (2) because terms a_{-j} decrease as $j \to -\infty$, so that beyond a certain number $j = -s$ all terms can be neglected without significant loss of accuracy. We must clarify that in our perspective the number of nonnegative terms $s + 1$ is larger, by orders of magnitude, than p or q typically used in ARMA(p, q) models. Also, the number s is totally unrelated to the number of essential parameters $m + 2$ of the autocovariance function, discussed in section 2, as coefficients a_j are internal parameters of the computational scheme. By contrast, the number s could be regarded as a large number of the order of magnitude 100 or 1000, depending on the decay of autocovariance, the desired accuracy, and the simulation length. In this respect, (18) and (19) can be approximated by

$$X_i = \sum_{j=-s}^{0} a_{-j} V_{i+j} = a_s V_{i-s} + \cdots + a_2 V_{i-2} + a_1 V_{i-1} + a_0 V_i, \qquad (20)$$

$$\sum_{j=0}^{s-i} a_j a_{i+j} = \gamma_i \qquad i = 0, 1, 2, \ldots, \qquad (21)$$

respectively, for a sufficiently large s.

Extending this notion, we can write X_i as the weighted sum of both previous and next (theoretically infinite) innovation variables V_i, i.e., in the following backward-forward moving average (BFMA) form:

$$X_i = \sum_{j=-\infty}^{\infty} a_j V_{i+j} = \cdots + a_{-1} V_{i-1} + a_0 V_i + a_1 V_{i+1} + \cdots , \tag{22}$$

where now the coefficients a_j are related to γ_j through the almost obvious equation

$$\sum_{j=-\infty}^{\infty} a_j a_{i+j} = \gamma_i \qquad i = 0, 1, 2, \ldots, n. \tag{23}$$

In section 5 we will see that the introduction of forward innovation terms (i.e., V_{i+1}, V_{i+2}, etc.) does not create any inconvenience even if the model is going to be used as a forecast model.

The backward-forward moving average model (22) is more flexible than the typical backward moving average model (18). Indeed, the number of parameters a_j in model (22) is double that of model (18) in order to represent the same number of autocovariances γ_j. Therefore in model (22) there exists an infinite number of sequences a_j satisfying (23).

One of the infinite solutions of (23) is that with $a_j = 0$ for every $j < 0$, in which case the model (22) is identical to the model (18). Another interesting special case of (22) is that with

$$a_j = a_{-j} \qquad j = 1, 2, \ldots \tag{24}$$

For reasons that will be explained below, the latter case will be adopted as the preferable model throughout this paper and will be referred to as the symmetric moving average (SMA) model (although the BMA model will be considered as well). In this case, (22) can be written as

$$X_i = \sum_{j=-s}^{s} a_{|j|} V_{i+j} = a_s V_{i-s} + \cdots + a_1 V_{i-1} + a_0 V_i + a_1 V_{i+1} + \cdots + a_s V_{i+s}, \tag{25}$$

where we have also restricted the number of innovation variables to a finite number, for the practical reasons already explained above. That is, we have assumed $a_j = 0$ for $|j| > s$. The equations relating the coefficients a_j to γ_j become now

$$\sum_{j=-s}^{s-i} a_{|j|} a_{|i+j|} = \gamma_i \qquad i = 0, 1, 2, \ldots \tag{26}$$

Given that the internal model parameters a_j are $s + 1$ in total, the model can preserve the first $s + 1$ terms of the autocovariance γ_j of the process X_i, if a_j are calculated so that (26) is satisfied for $i = 0, \ldots, s$. (In section 4 we will discuss how this calculation can be done.) As we have already discussed, the number s can be chosen so that the desired accuracy can be achieved. The model implies nonzero autocovariance for a number of subsequent time lags. Thus for the subsequent s terms ($j = s + 1, \ldots, 2s$) the autocovariance terms are given by

$$\gamma_i = \sum_{j=i-s}^{s} a_j a_{i-j} \qquad i = s + 1, \ldots, 2s \tag{27}$$

(a consequence of (26) for $i > s$), whereas for even larger lags the autocovariance vanishes off.

Apart from the parameters a_j that are related to the autocovariance of the process X_i, two more parameters are needed for the generating scheme, which are related to the mean and skewness of the process. These are the mean $\mu_V := E[V_i]$ and the coefficient of skewness $\xi_V := E[(V_i - \mu_V)^3]$ of the innovations V_i (note that by definition, Var $[V_i] = 1$). They are related to the corresponding parameters of X_i by

$$\left(\sum_{j=0}^{s} a_j \right) \mu_V = \mu_X, \qquad \left(\sum_{j=0}^{s} a_j^3 \right) \xi_V = \xi_X \gamma_0^{3/2} \tag{28}$$

for the BMA model and

$$\left(a_0 + 2 \sum_{j=1}^{s} a_j \right) \mu_V = \mu_X, \qquad \left(a_0^3 + 2 \sum_{j=0}^{s} a_j^3 \right) \xi_V = \xi_X \gamma_0^{3/2} \tag{29}$$

for the SMA model, which are direct consequences of (20) and (25), respectively.

To provide a more practical view of the behavior of the SMA model, also in comparison with the typical BMA model, we demonstrate in Figure 2 two examples in graphical form. In the first example, we have assumed that the process X_i is Markovian with autocovariance (3) and $\gamma_0 = 1$ and $\rho = 0.9$. In case of the BMA model with infinite a_j terms, a theoretical solution of (19) is

$$a_j = \sqrt{\gamma_0 (1 - \rho^2)} \rho^j, \tag{30}$$

as can be easily verified by substituting (30) into (19). If we choose to preserve the first 101 autocovariance terms γ_j assuming that $a_j = 0$ for $j > s = 100$, we can numerically estimate from (21) the first 101 nonzero terms a_j (in a manner that will be described in section 4). The numerically estimated a_j are depicted in Figure 2a; practically, they equal those given by (30), apart from the last three values, which depart from theoretical values because of the effect of setting the high terms $a_j = 0$ (the departure is clear in Figure 2a). The same autocovariance γ_j can be also preserved by the SMA model. The theoretical solution for infinite a_j terms, and the approximate solution, again using 101 nonzero a_j terms, are calculated from (23) and (26), respectively (using techniques that will be described in section 4), and are also shown in Figure 2a. We observe that all a_j values of the SMA model (apart from a_0) are smaller than the corresponding a_j values of the BMA model; for large j near 100, a_j of the SMA model become 1 order of magnitude smaller than those of the BMA model. Apparently, this constitutes a strong advantage of the SMA model over the BMA one: The smaller the coefficients a_j for large j, the smaller is the introduced error due to setting $a_j = 0$ for $j > s$.

In a second example we have assumed that X_i is a FGN process with autocovariance (5), and $\gamma_0 = 1$ and $H = 0.6$, which corresponds to $\beta = 1.25$. This autocovariance is shown graphically in Figure 2b along with the resulting sequences of a_j, assuming again that the first 101 terms are nonzero. Once more we observe that the a_j sequence of the SMA model lies below that of the BMA model. In addition to the approximate solution for 101 nonzero a_j terms a theoretical solution for infinite a_j terms, also shown in Figure 2b, is possible for the

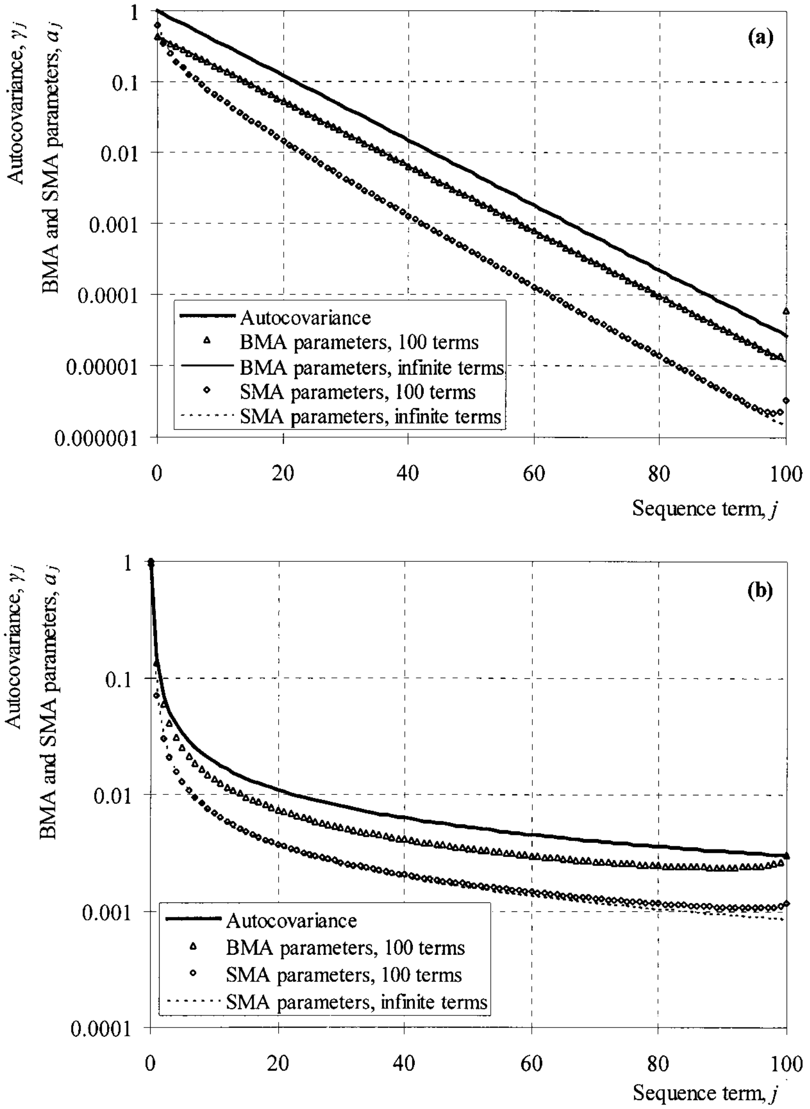

Figure 2. Two examples of theoretical autocovariance sequences and resulting sequences of internal parameters a_j for BMA and SMA schemes: (a) a Markovian autocovariance sequence and (b) a fractional Gaussian noise autocovariance sequence. The obtained autocovariance sequences by either of the BMA or SMA schemes are indistinguishable from the theoretical ones.

SMA model, as will be described in section 4.1. We will also see in section 4.1 and Appendix A2 that a closed analytical solution is possible for the SMA model for any autocovariance γ_j but not for the BMA model. This certainly constitutes a second advantage of the SMA model over the BMA one.

As we have already mentioned above, the SMA model implies a nonzero autocovariance even for lags above the assumed numerical limit s, i.e., for $j = s + 1$ up to $j = 2s$, given by (27). On the contrary, the BMA model implies that all autocovariance terms above s are zero. In Figure 3 we have plotted the resulting autocovariance of the above-described Markovian example for lags j up to 200. We observe that this structure may be an accepted approximation of the Markovian structure for lags 101–200 (at least it is better than the zero autocovariance implied by the BMA model). As this is achieved by no cost at all (no additional parameters are introduced), it can be regarded as an additional advantage of the SMA model over the BMA model.

A fourth advantage of the SMA model is related to the preservation of skewness, in cases of skewed variables, which are very common in stochastic hydrology. It is well known [*Todini*, 1980; *Koutsoyiannis*, 1999a] that if the coefficient of skewness of the innovation variables becomes too high, it is impossible to preserve the skewness of the variables X_i. Therefore the model resulting in lower coefficient of skewness of the innovation variables is preferable. In all cases examined, this was the SMA model. For instance, in the above-described Markovian example the SMA model resulted in $\xi_V = 2.52\ \xi_X$

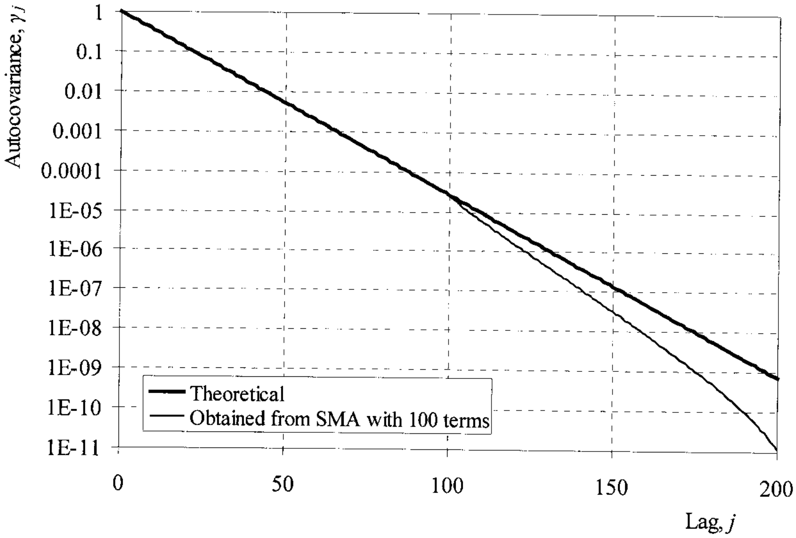

Figure 3. Obtained autocovariance structure from the SMA scheme using 100 a_j terms, for lags 0–200; the theoretical autocovariance structure is that of Figure 2a.

whereas, in the BMA model, $\xi_V = 3.27\ \xi_X$ (by applying (29) and (28), respectively).

4. Computation of Internal Parameters of the Generating Scheme

We will present two methods for computing the sequence of terms a_j given the autocovariance γ_j. The first method results in closed analytical solution of (23) for the case that (24) holds; this is applicable to the SMA model for an infinite number of a_j terms. The second method is a numerical solution of (21) or (26) that determines a finite number of a_j terms and is applicable to both the BMA and the SMA models.

4.1. Closed Solution

Denoting $s_a(\omega)$ the DFT of the a_j series and utilizing the convolution equation (23) and the fact that in the SMA model, a_j is an even function of j (equation (24)), we can show (see Appendix A2) that $s_a(\omega)$ is related to the power spectrum of the process $s_\gamma(\omega)$ by

$$s_a(\omega) = \sqrt{2s_\gamma(\omega)}. \tag{31}$$

This enables the direct calculation of the DFT of the a_j series if the power spectrum of the process $s_\gamma(\omega)$ (or equivalently, the autocovariance γ_j) is known. Then a_j can be calculated by the inverse transform, i.e.,

$$a_j = \int_0^{1/2} s_a(\omega) \cos(2\pi j\omega)\, d\omega. \tag{32}$$

Apart from a few special cases, the calculations needed to evaluate a_j from γ_j can be performed only numerically. However, they are simple and noniterative. In addition, all calculations can be performed using the fast Fourier transform (FFT, e.g., *Bloomfield* [1976, pp. 61–76]), thus enabling the building of a fast algorithm.

For the BMA model the fact that a_j is not an even (or an odd) function of j results in a complex DFT of a_j. Therefore the corresponding relation between $s_a(\omega)$ and $s_\gamma(\omega)$ becomes (see Appendix A2)

$$|s_a(\omega)| = \sqrt{2s_\gamma(\omega)}, \tag{33}$$

where $|s_a(\omega)|$ is the absolute value of $s_a(\omega)$. Given that $s_a(\omega)$ is complex, (33) does not suffice to calculate $s_a(\omega)$ (it gives only its amplitude, not its phase). Therefore the method cannot work for the BMA model. In addition, it is shown in Appendix A2 that there does not exist any other real valued transformation, different from DFT, that could result in an equation similar to (31) to enable a direct calculation of a_j for the BMA model. However, the iterative method presented in section 4.2 can be applied to both the SMA and the BMA models.

4.2. Iterative Solution

The equations relating the model internal parameters a_j to the autocovariance terms γ_j, i.e., (21) and (26) for the BMA and SMA models, respectively, may be written simultaneously for $j = 0, \ldots, s$ in matrix notation as

$$\mathbf{p}\boldsymbol{\zeta} = \boldsymbol{\theta}, \tag{34}$$

where $\boldsymbol{\zeta} = [a_0, \ldots, a_s]^T$, $\boldsymbol{\theta} = [\gamma_0, \ldots, \gamma_s]^T$ (with the exponent T denoting the transpose of a matrix or vector), and $\mathbf{p}$ is a matrix with size $(s + 1) \times (s + 1)$ and elements

$$p_{ij} = (1/2)[a_{j-i}U(j - i) + a_{i+j-2}U(s - i - j + 1)] \tag{35}$$

for the BMA model and

$$p_{ij} = a_{|j-i|} + a_{i+j-2}U(j - 2)U(s - i - j + 1) \tag{36}$$

for the SMA model. Here $U(x)$ is the Heaviside's unit step function, with $U(x) = 1$ for $x \geq 0$ and $U(x) = 0$ for $x < 0$. It can be easily verified that (35) and (36) (along with (34)) are equivalent to (21) and (26), respectively. Other expressions equivalent to (35) and (36) and simpler than them can be also derived, but (35) and (36) are the most convenient in subsequent steps.

Clearly, each single equation of the system (34) includes second-order products of unknown terms a_j. Therefore (34) may have one or more solutions in case of a positive definite autocovariance or no solution otherwise. Generally, we need to

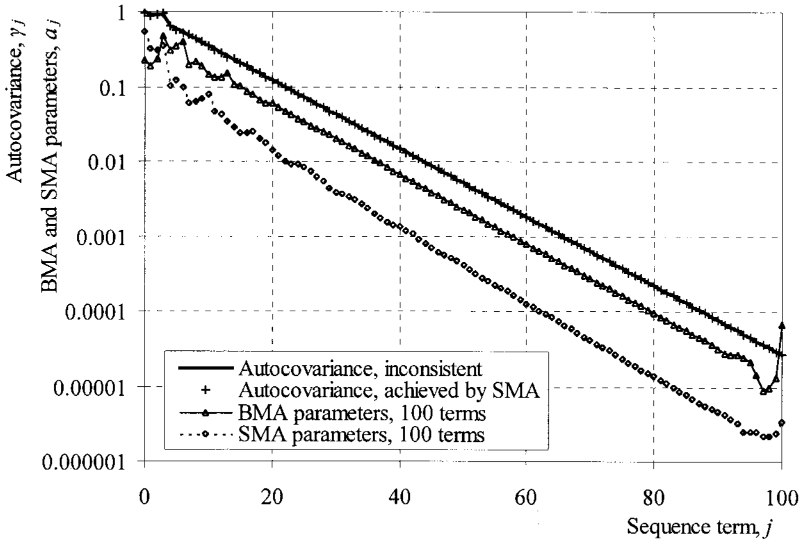

Figure 4. An example of an inconsistent γ_j sequence approximated with a consistent sequence achieved by the SMA scheme using 100 a_j terms; the latter are also plotted, in comparison with the corresponding terms of the BMA scheme.

determine one single solution if it exists; otherwise, we may need to find the best approximation to (34). To accomplish these in a common manner, we reformulate the parameter estimation problem as a minimization problem, demanding to

$$\min f(\boldsymbol{\zeta}) = f(a_0, \ldots, a_s) := \|\mathbf{p}\boldsymbol{\zeta} - \boldsymbol{\theta}\|^2 + \lambda(\mathbf{p}_1\boldsymbol{\zeta} - \gamma_0)^2, \tag{37}$$

where $\mathbf{p}_1$ is the first row of $\mathbf{p}$, λ is a weighting factor, and $\|.\|$ denotes the Euclidean norm of a vector. The meaning of the first term of the right-hand side of (37) becomes obvious from (34). The second term denotes the square error in preserving the model variance γ_0, multiplied by the weighting factor λ. Although, apparently, the second term is also contained in the first term, its separate appearance in the objective function enables its separate treatment. In case of a feasible autocovariance sequence, the minimum of $f(\boldsymbol{\zeta})$ will be zero, whatever the value of λ. However, in case of an inconsistent autocovariance sequence the minimum of $f(\boldsymbol{\zeta})$ will be a positive number. In such a case the preservation of the variance γ_0 is more important than that of autocovariance terms. Assigning a large value to λ (e.g., $\lambda = 10^3$), we force $(\mathbf{p}_1\boldsymbol{\zeta} - \gamma_0)^2$ to take a value close to zero. Alternatively, λ could be considered as a Lagrange multiplier (an extra variable of the objective function (37)), but this would complicate the solution procedure.

The task of minimization of $f(\boldsymbol{\zeta})$ is facilitated by determining its derivatives with respect to $\boldsymbol{\zeta}$. After algebraic manipulations it can be shown that $d(\mathbf{p}\boldsymbol{\zeta})/d\boldsymbol{\zeta} = 2\mathbf{p}$ (for both BMA and SMA schemes), so that

$$\frac{df(\boldsymbol{\zeta})}{d\boldsymbol{\zeta}} = 4(\mathbf{p}\boldsymbol{\zeta} - \boldsymbol{\theta})^T\mathbf{p} + 4\lambda(\mathbf{p}_1\boldsymbol{\zeta} - \gamma_0)\mathbf{p}_1. \tag{38}$$

Clearly, the problem we have to solve is an unconstrained nonlinear optimization problem with analytically determined derivatives. This can be easily tackled by typical methods of the literature such as the steepest descent and Fletcher-Reeves conjugate gradient methods [e.g., *Mays and Tung*, 1996, p. 6.12; *Press et al.*, 1992, p. 422]. These are iterative methods, starting with an initial vector, which, in our case, can be taken as $\boldsymbol{\zeta}^{[0]} = [\sqrt{\gamma_0}, 0, 0, \ldots, 0]^T$, and iteratively improving it until the solution converges.

The algorithm has been proven very quick and efficient in all cases examined, involving problems even with more than 1000 a_i parameters. Examples of applying the algorithm for consistent autocovariances, Markovian and fractional Gaussian, have been already discussed (section 2 and Figure 2). An example of applying the algorithm to an inconsistent autocovariance is shown in Figure 4. The autocovariance of this example is identical to that of the Markovian example of Figure 2a, apart from the values γ_2 and γ_3 that were both set equal to γ_1; this creates a covariance matrix $\mathbf{h}$ that is not positive definite. As shown in Figure 4, the algorithm resulted in a very good approximation of the assumed autocovariance.

In comparison with an earlier numerical procedure by *Wilson* [1969; see also *Box and Jenkins*, 1970, p. 203] for determining the parameters of the BMA process, the above-described algorithm is more general (it also covers the SMA case), faster (it does not involve matrix inversion, whereas Wilson's algorithm does), and more flexible and efficient (it can provide approximate solutions for inconsistent autocovariances, whereas Wilson's algorithm cannot).

5. Generation Scheme in Forecast Mode

Equations (20) and (25) are directly applicable for simulation (unconditional generation) of the process X_i. However, it is quite frequently the case where some of the variables X_i (past and present) are known and we wish either to generate other (future) variables, or to obtain best predictions of these (future) variables. As we will see, both problems can be tackled in a common simple manner, applicable for both the BMA and SMA models.

We will assume that the vector consisting of the present and k past variables $\mathbf{Z} := [X_0, X_{-1}, \ldots, X_{-k}]^T$ is known and its value is $\mathbf{z} = [x_0, x_{-1}, \ldots, x_{-k}]^T$. We wish either to generate any future variable X_j for $j > 0$, or to predict its value, under the condition $\mathbf{Z} = \mathbf{z}$. These can be done utilizing the following proposition, whose proof is given in Appendix A3:

Proposition. Let $\tilde{X}_i (i = -k, \ldots, 0, 1, 2, \ldots)$ be any discrete time stochastic process with autocovariance $\gamma_j (j = 0, 1, \ldots)$ and let $\tilde{\mathbf{Z}} := [\tilde{X}_0, \tilde{X}_{-1}, \ldots, \tilde{X}_{-k}]^T$. Let also $\mathbf{Z} := [X_0, X_{-1}, \ldots, X_{-k}]^T$ be a vector of stochastic variables independent of $\tilde{X}_i$ with mean and autocovariance identical to that of $\tilde{X}_i$. Then, the stochastic process

$$X_i = \tilde{X}_i + \boldsymbol{\eta}_i^T \mathbf{h}^{-1} (\mathbf{Z} - \tilde{\mathbf{Z}}) \qquad i = 1, 2, \ldots, \tag{39}$$

where $\boldsymbol{\eta}_i^T :=$ Cov $[\tilde{X}_i, \tilde{\mathbf{Z}}]$ and $\mathbf{h} :=$ Cov $[\tilde{\mathbf{Z}}, \tilde{\mathbf{Z}}]$, has identical mean and autocovariance with those of $\tilde{X}_i$. In addition, the conditional variance of X_i, given $\mathbf{Z} = \mathbf{z}$, is

$$\text{Var } [X_i | \mathbf{Z} = \mathbf{z}] = \gamma_0 - \boldsymbol{\eta}_i^T \mathbf{h}^{-1} \boldsymbol{\eta}_i \tag{40}$$

and is identical to the least mean squares prediction error of X_i from $\mathbf{Z}$.

Note that $\mathbf{h}$ is a symmetric matrix with size $(k + 1) \times (k + 1)$ and elements given by (12) whereas $\boldsymbol{\eta}_i$ is a vector with size $k + 1$ and elements

$$(\boldsymbol{\eta}_i)_j = \gamma_{|i+j-1|}. \tag{41}$$

Also, note that the proposition is quite general and it can be applied to any type of linear stochastic model (not only to those examined in this paper).

This proposition enables the following procedure for the forecast mode of the model: (1) Determine the matrix $\mathbf{h}$ using (12) for the given number $(k + 1)$ of known (present and past) variables and then calculate $\mathbf{h}^{-1}$. (2) Generate a sequence of variables $\tilde{X}_i (i = -k, \ldots, 0, 1, \ldots)$, using the adopted model (20) or (25) without any reference to the known variables $\mathbf{Z}$. Form the vectors $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ and calculate the vector $\mathbf{h}^{-1}(\mathbf{Z} - \tilde{\mathbf{Z}})$. (3) For each $i > 0$, determine the vector $\boldsymbol{\eta}_i$ from (41) and calculate the final value of the variable X_i, conditional on $\mathbf{Z}$, from (39).

Equation (40) shows that the conditional variance of X_i is smaller than the unconditional one (γ_0), as expected. The fact that this conditional variance is identical to the least mean squares prediction error of X_i from $\mathbf{Z}$ ensures us that no further reduction is possible by any type of linear prediction model. Thus the combination of model (20) or (25) with the transformation (39) allows preservation of the stochastic structure of the process, whatever this structure is, and simultaneously reduces the conditional variance to its smallest possible value, in the sense that no other linear stochastic model could reduce it further. Notably, the same generating model (20) or (25) is used in both modes, simulation and forecast.

Theoretically, the procedure can be applied to negative values of i, as well. In this case, if $-k \le i \le 0$, it is easy to show that (39) reduces to the trivial case $X_i = X_i$, as it should (see Appendix A3).

The above steps are appropriate if the forecast is done in terms of conditional simulation. If it must be done in terms of expected values rather than conditionally simulated values, then in step 2 of the above procedure, $\tilde{X}_i$ are set equal to their (unconditional) expected values rather than generated. In this case, if confidence limits are needed, they can be calculated in terms of the conditional variance given by (40).

6. Multivariate Case

The model studied in sections 2–5 is a single-variate model but can be easily extended to the multivariate case. In this case the model, apart from the temporal covariance structure, should consider and preserve the contemporaneous covariance structure of several variables corresponding to different locations.

Let $\mathbf{X}_i = [X_i^1, X_i^2, \ldots, X_i^n]^T$ be the vector of n stochastic variables each corresponding to some location specified by the index $l = 1, \ldots, n$, at a specific time period i. Let also $\mathbf{g}$ be the variance-covariance matrix of those variables with elements

$$g^{lk} := \text{Cov } [X_i^l, X_i^k] \qquad l, k = 1, 2, \ldots, n. \tag{42}$$

We assume that each of the variables X_i^l can be expressed in terms of some auxiliary variables V_i^l (again with unit variance) by using either

$$X_i^l = \sum_{r=-s}^{0} a_{-r}^l V_{i+r}^l \tag{43}$$

for the BMA model or

$$X_i^l = \sum_{r=-s}^{s} a_{|r|}^l V_{i+r}^l \tag{44}$$

for the SMA model. These equations are similar to (20) and (25), respectively.

The auxiliary variables V_i^l can be assumed uncorrelated in time i (i.e., Cov $[V_i^l, V_m^k] = 0$ if $i \neq m$) but correlated in different locations l for the same time i. If $\mathbf{c}$ is the variance-covariance matrix of variables V_i^l, then each of its elements

$$c^{lk} := \text{Cov } [V_i^l, V_i^k] \qquad l, k = 1, 2, \ldots, n \tag{45}$$

can be expressed in terms of g^{lk} and the series of a_i^l and a_i^k by

$$c^{lk} = g^{lk} \Big/ \sum_{r=0}^{s} a_r^l \, a_r^k \tag{46}$$

for the BMA model and

$$c^{lk} = g^{lk} \Big/ \sum_{r=-s}^{s} a_{|r|}^l \, a_{|r|}^k \tag{47}$$

for the SMA model. These equations are direct consequences of (43) and (44), respectively. The theoretically anticipated lagged cross-covariance for any lag $j = 0, 1, \ldots,$ is then

$$\text{Cov } [X_i^l, X_{i+j}^k] = g^{lk} \sum_{r=0}^{s-j} a_r^l \, a_{j+r}^k \Big/ \sum_{r=0}^{s} a_r^l \, a_r^k \tag{48}$$

for the BMA model and

$$\text{Cov } [X_i^l, X_{i+j}^k] = g^{lk} \sum_{r=-s}^{s-j} a_{|j+r|}^l a_{|r|}^k \Big/ \sum_{r=-s}^{s} a_{|r|}^l a_{|r|}^k \tag{49}$$

for the SMA model.

Given the variance-covariance matrix $\mathbf{c}$, the vector of variables $\mathbf{V}_i = [V_i^1, V_i^2, \ldots, V_i^n]^T$ can be generated using the simple multivariate model

$$\mathbf{V}_i = \mathbf{b}\mathbf{W}_i, \tag{50}$$

where $\mathbf{W}_i = [W_i^1, W_i^2, \ldots, W_i^n]^T$ is a vector with innovation variables with unit variance independent both in time i and in

Table 1. Theoretical and Empirical Statistics of the Application of Section 6

	Theoretical		Empirical	
	Location 1	Location 2	Location 1	Location 2
Mean	1.00	2.00	1.00	1.97
Standard deviation	0.50	1.20	0.51	1.21
Coefficient of skewness	1.00	1.20	1.03	1.14
Hurst coefficient	0.60	0.70	0.61	0.71
Cross-correlation coefficient	0.70		0.70	

location $l = 1, \ldots, n$ and $\mathbf{b}$ is a matrix with size $n \times n$ such that

$$\mathbf{b}\mathbf{b}^T = \mathbf{c}. \qquad (51)$$

The methodology for solving (51) for $\mathbf{b}$ given $\mathbf{c}$ (also known as taking the square root of $\mathbf{c}$) will be discussed in section 7 below. The other parameters needed to completely define model (50) are the vector of mean values $\boldsymbol{\mu}_\mathbf{W}$ and coefficients of skewness $\boldsymbol{\xi}_\mathbf{W}$ of W_i^l. These can be calculated in terms of the corresponding vectors $\boldsymbol{\mu}_\mathbf{V}$ and $\boldsymbol{\xi}_\mathbf{V}$ of V_i^l, already known from (28) or (29), by

$$\boldsymbol{\mu}_\mathbf{W} = \mathbf{b}^{-1}\boldsymbol{\mu}_\mathbf{V}, \qquad \boldsymbol{\xi}_\mathbf{W} = (\mathbf{b}^{(3)})^{-1}\boldsymbol{\xi}_\mathbf{V}, \qquad (52)$$

which are direct consequences of (50). In (52), $\mathbf{b}^{(3)}$ is the matrix whose elements are the cubes of $\mathbf{b}$, and the exponent -1 denotes the inverse of a matrix.

To illustrate the method, we have applied it to a problem with two locations with statistics given in Table 1. To investigate the method's ability to preserve long-term memory properties such as the Hurst coefficient in multiple dimensions, we have assumed the FGN structure with exponents β equal to 1.25 and 1.667 for locations 1 and 2, respectively, corresponding to Hurst coefficients 0.6 and 0.7 for locations 1 and 2, respectively. We generated a synthetic record with 10,000 data values using the SMA scheme with 2000 nonzero a_j terms, which were evaluated by the closed solution described in section 4.1. The last (2000th) term of the series of a_j was $6 \times 10^{-5}\, a_0$ for location 1 and $3 \times 10^{-4}\, a_0$ for location 2; these small values indicate that the error due to neglecting the higher a_j terms (beyond term 2000) is small. The required computer time on a modest (300 MHz) Pentium PC was ~10 s for the computation of internal parameters (when the fast Fourier transform was implemented in the algorithm; otherwise it increased to ~2 min) and another 10 s for the generation of the synthetic records. As shown in Table 1, the preservation of all statistics was perfect. In addition, Figure 5 shows that the autocorrelation and cross-correlation function, the power spectrum, and the rescaled range as a function of record length were very well preserved, as well.

7. Finite Length of Autocorrelation Sequence

In sections 1–6 it was assumed that the autocovariance γ_j is defined for any arbitrarily high lag j. However, there are cases where only a finite number of autocovariance terms can be defined. For example, in a stochastic model describing rainfall increments at time intervals δ within a rainfall event with certain duration $d = q\delta$ (where q is an integer), the autocovariance has no meaning for lags greater than $q - 1$ (see the application in the end of this section). Such cases can be tackled in a different, rather simpler, way.

An appropriate model for this case is

$$\mathbf{X} = \mathbf{b}\mathbf{V}, \qquad (53)$$

where $\mathbf{X} = [X_1, \ldots, X_q]^T$ is the vector of variables to be modeled with variance-covariance matrix $\mathbf{h}$ given by (12), $\mathbf{V} = [V_1, \ldots, V_q]^T$ is a vector of innovations with unit variance, and $\mathbf{b}$ is a square matrix of coefficients with size $q \times q$. The main difference from the models of sections 3–5 is that the number of innovations $\mathbf{V}$ equals the number q of the modeled variables $\mathbf{X}$ (the length of the synthetic record). In this case the distributions of innovations $\mathbf{V}$ cannot be identical. Each one has different mean and coefficient of skewness, given by

$$\boldsymbol{\mu}_\mathbf{V} = \mathbf{b}^{-1}\boldsymbol{\mu}_\mathbf{X}, \qquad \boldsymbol{\xi}_\mathbf{V} = (\mathbf{b}^{(3)})^{-1}\boldsymbol{\xi}_\mathbf{X}, \qquad (54)$$

which are direct consequences of (53). The matrix of coefficients $\mathbf{b}$ is given by

$$\mathbf{b}\mathbf{b}^T = \mathbf{h}, \qquad (55)$$

which again is a direct consequence of (53).

It is reminded that (55) has an infinite number of solutions $\mathbf{b}$ if $\mathbf{h}$ is positive definite. Traditionally, two well-known algorithms are used which result in two different solutions $\mathbf{b}$ [see, e.g., *Bras and Rodriguez-Iturbe*, 1985, p. 96; *Koutsoyiannis*, 1999a]. The first and simpler algorithm, known as triangular or Cholesky decomposition, results in a lower triangular $\mathbf{b}$. The second, known as singular value decomposition, results in a full $\mathbf{b}$ using the eigenvalues and eigenvectors of $\mathbf{h}$. A third algorithm has been proposed by *Koutsoyiannis* [1999a] which is based on an optimization framework and can determine any number of solutions, depending on the objective set (for example, the minimization of skewness, or the best approximation of the covariance matrix, in case that it is not positive definite).

We can observe that the lower triangular $\mathbf{b}$ is directly associated with the BMA model discussed in section 3, but with different number of innovations V_i for each X_i. Thus, if $\mathbf{b}$ is lower triangular, then, apparently, $X_1 = b_{11}V_1$, $X_2 = b_{21}V_1 + b_{22}V_2$, etc. Likewise, a symmetric $\mathbf{b}$ is associated with the SMA model. An iterative method for deriving a symmetric $\mathbf{b}$ can be formulated as a special case of the methodology proposed by *Koutsoyiannis* [1999a]. This can be based on the minimization of

$$f(\mathbf{b}) := \|\mathbf{b}\mathbf{b}^T - \mathbf{h}\|^2, \qquad (56)$$

where we have used the notation $\|\mathbf{b}\mathbf{b}^T - \mathbf{h}\|$ for the norm (more specifically, we adopt the Euclidean or standard norm here; see, e.g., *Marlow* [1993, p. 59]), as if $\mathbf{b}\mathbf{b}^T - \mathbf{h}$ were a vector in q^2 space rather than a matrix. The derivatives of $f(\mathbf{b})$ with respect to $\mathbf{b}$ are easy to determine (see Appendix A4). Using the notation $d\alpha/d\mathbf{b} = [\partial\alpha/\partial b_{ij}]$ for the matrix of partial derivatives of any scalar a with respect to all b_{ij} (this is an extension of the notation used for vectors, e.g., *Marlow* [1993, p. 208]) and considering that $\mathbf{b}$ is symmetric, we find that

$$\frac{df(\mathbf{b})}{d\mathbf{b}} = 8\mathbf{e} - 4\mathbf{e}^*, \qquad (57)$$

where $\mathbf{e} := (\mathbf{b}\mathbf{b}^T - \mathbf{h})\mathbf{b}$ and $\mathbf{e}^* = \operatorname{diag}(e_{11}, e_{22}, \ldots, e_{qq})$, i.e., a diagonal matrix containing the diagonal elements of $\mathbf{e}$.

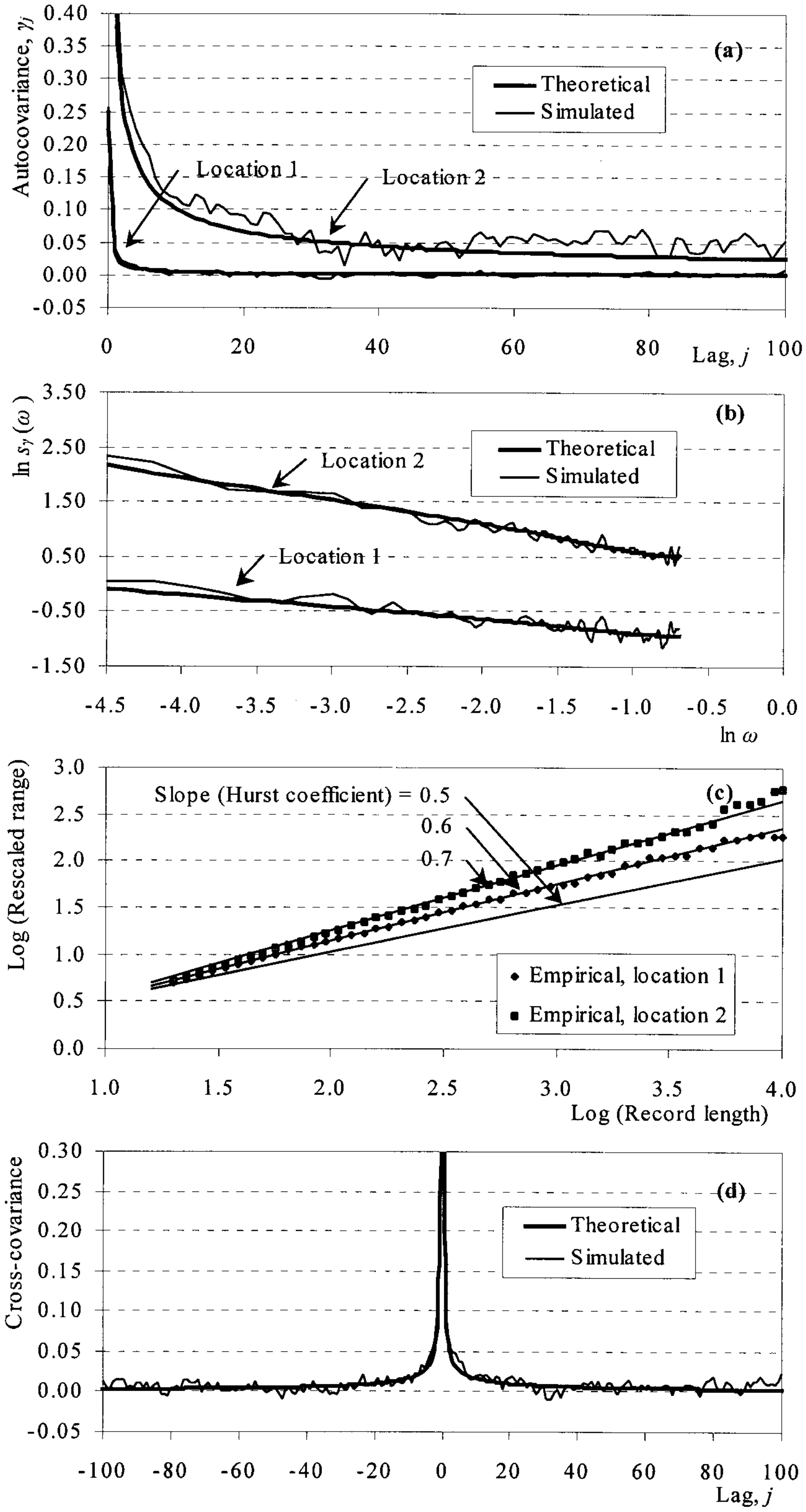

Figure 5. Preservation of statistical properties by the simulated records of the application of section 6: (a) autocovariance, (b) power spectra, (c) rescaled range and Hurst coefficients, and (d) cross-covariance.

As in the similar case of section 4.2, the problem here is an unconstrained nonlinear optimization problem with analytically determined derivatives, which can be easily tackled by typical methods such as the steepest descent and the Fletcher-Reeves conjugate gradient methods. As initial solution for the iterative procedure we should use a symmetric one; a good choice is $\mathbf{b}^{[0]} = \sqrt{\gamma_0}\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

To illustrate the method and the differences among the three different solutions discussed, we have considered a stochastic model of a rainfall event with duration $d = 20$ hours using a half-hour time resolution δ, so that the number of variables is $q = 20/0.5 = 40$. We denote $X_i (i = 1, \ldots, 40)$ the half-hour rainfall increments and assume that the covariance structure of X_i is as in the scaling model of storm hyetograph [*Koutsoyiannis and Foufoula-Georgiou*, 1993]; that is,

$$\gamma_{|i-j|} = \mathrm{Cov}\,[X_i, X_j] = [(c_2 + c_1^2)\varphi(|j - i|, \beta)q^{1/\beta} - c_1^2] \cdot (d^{2(\kappa+1)}/q^2), \qquad (58)$$

where c_1, c_2, κ and β are parameters and

$$\varphi(m, \beta) := (1/2)[(m - 1)^{2-1/\beta} + (m + 1)^{2-1/\beta}] - m^{2-1/\beta} \qquad m > 0, \qquad (59)$$

whereas $\varphi(0, \beta) = 1$. This is apparently a long-memory autocorrelation structure similar to the FGN structure. It always results in consistent (positive definite) autocovariance if it is evaluated within the duration d of the event; however, for certain combinations of parameters it can result in inconsistent autocovariance values if it is attempted to evaluate it outside of the event (i.e., for lags greater than $q - 1$).

For the example presented here we have assumed that the model parameters are $c_1 = 8.74$, $c_2 = 85.68$, $\kappa = -0.45$, and $\beta = 10$ (units of millimeters and hours). The statistics of X_i, determined from equations given by *Koutsoyiannis and Foufoula-Georgiou* [1993], are $\mu_X = 1.14$ mm, $\gamma_0 = 2.68$ mm^2, and $\xi_X = 2.88$ (the latter is determined assuming two-parameter gamma distribution for X_i). The matrix **b** is 40×40 (1600 elements). We have calculated all three solutions of the matrix **b** described above (triangular, singular value, and symmetric), which are shown schematically in Figure 6. We observe a regular pattern with a strong diagonal and a strong first column for the triangular solution, a strong first column and an irregular pattern for other columns for the singular value solution, and a regular pattern with a strong diagonal for the symmetric solution.

An appropriate means to compare the three solutions is provided by the resulting coefficients of skewness of innovations V_i, given by (54). These are shown in Figure 7. The singular value solution resulted in coefficients of skewness ranging from −40 to +62, which apparently are computationally intractable at generation. More reasonable are the values of the triangular solution, with a maximum coefficient of skewness equal to ~10. The symmetric solution resulted in the smallest, among the three cases, maximum coefficient of skewness, slightly exceeding 6. Notably, this value is the smallest possible value among all possible (infinite) **b** solutions of (55) [*Koutsoyiannis*, 1999b]. This enhances further the already discussed feature of the SMA model, that symmetric solutions result in smaller coefficients of skewness of innovations, a feature quite expedient in stochastic hydrology.

The finite length scheme described in this section can be a preferable alternative even in cases where the autocovariance is defined for any j but the length q of the synthetic record is very small. Specifically, the scheme of the present section 7 uses q^2 internal parameters. In the case that the process exhibits long memory the required number of parameters s of the schemes of section 3 may be greater than q^2, and thus the scheme of section 7 could be preferable.

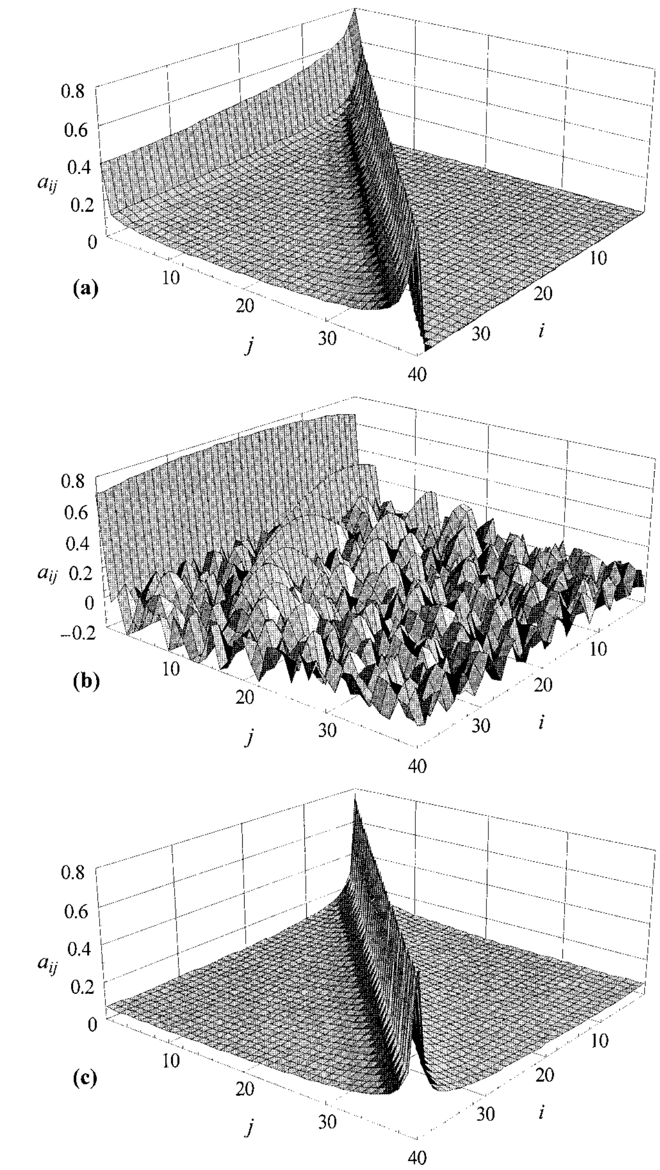

Figure 6. Comparison of three different solutions of parameter matrices **b** (3-D plots of their elements) of the application of section 7: (a) triangular solution, (b) singular value solution, and (c) symmetric solution.

8. Summary, Conclusions, and Discussion

The main topics of the proposed framework can be summarized in the following points: (1) A generalized autocovariance function is introduced which unifies in a simple mathematical expression both short-memory (ARMA) and long-memory (FGN) models, considering them as special instances in a parametrically defined continuum, more comprehensive than these classes of models. (2) A moving average stochastic generation scheme is proposed that can implement the generalized autocovariance function (or any other autocovariance function). In addition to the traditional backward moving average scheme, a new time-symmetric (backward-forward) moving average scheme is proposed. It is computationally more convenient and also results in better treatment of processes with skewed distributions. (3) Two methods of determining the internal parameters of the generating scheme are proposed. The first is a

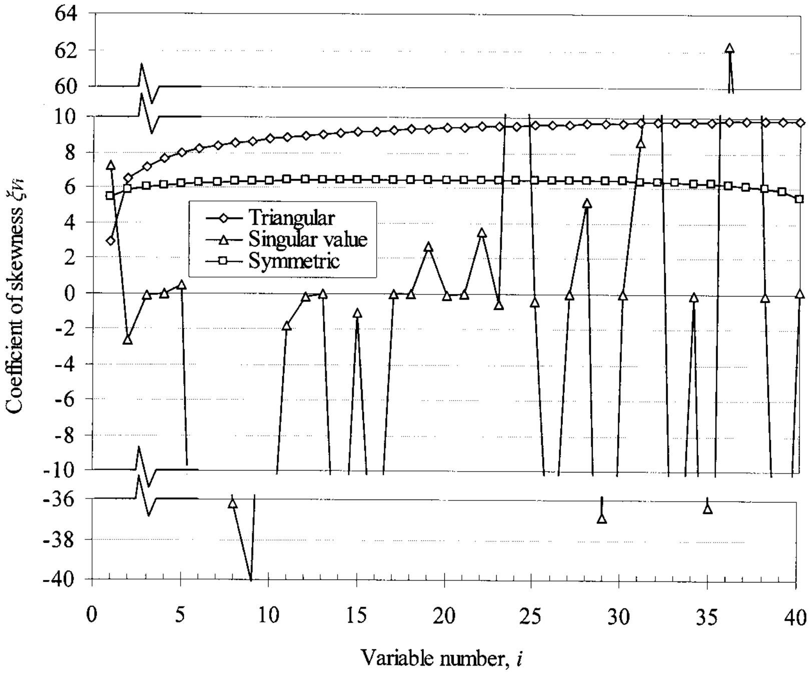

Figure 7. Comparison of the resulting coefficients of skewness of the 40 innovations of the application of section 7 for the three different solutions of parameter matrices **b**.

closed method based on the power spectrum of the process and applicable to the symmetric moving average scheme. The second is an iterative method based on convolution equations and applicable to both instances of the generating scheme. (4) The proposed stochastic generation scheme is directly adaptable so as to perform in forecast mode. To this aim, a generalized adaptation methodology is studied, applicable to any type of stochastic model. (5) The model can perform in single-variate as well as multivariate problems. (6) A specific form of the model for problems where the autocorrelation function can be defined only for a certain finite number of lags (e.g., in generation of rainfall increments within a rainfall event) is also studied. An incidental contribution of this study is a method for determining a symmetric square root of a symmetric matrix; this symmetric square root is the direct analogue of the symmetric moving average generating scheme, and, as demonstrated by an example, it outperformed nonsymmetrical solutions.

Thus the proposed framework is a generalized tool for any kind of single-variate and multivariate simulation and forecasting problems in stochastic hydrology involving stationary stochastic processes. We emphasize its appropriateness for modeling long-memory processes and its ability for preserving the Hurst coefficients in multivariate processes, even if each location has a different Hurst coefficient. Simultaneously, it enables explicit preservation of the skewness of the processes (at no computational or other cost, apart from generating skewed rather than Gaussian random numbers), a feature that is of major concern in stochastic hydrology. Owing to the proposed fast algorithms for computation of internal parameters the required computing time is small, even for problems including thousands of such parameters.

In traditional stochastic models, three different issues, i.e., the type of the generation scheme, the number of model parameters, and the type of autocovariance, are merged in one. For example, if we choose the AR(1) model as a generation scheme, we simultaneously choose to use two second-order parameters (variance and lag-one autocovariance) and assume that the autocovariance is an exponential function of the lag. In our approach, we have separated these three issues. The autocovariance function has a single mathematical expression of power type. The number of parameters can be decided separately, depending on the desired parsimony or nonparsimony of parameters and the length of the available record. The minimum number of parameters is three, one being the variance and another one the exponent of the power-type autocovariance function. This exponent equals zero for the model with the shortest possible memory, and becomes >1 for a long-memory model. Coming then to the generation scheme, this has a mathematical expression independent of the autocorrelation function. What we have to decide here is the number of innovation terms, which depends on the length of the synthetic record to be generated, the desired accuracy, and the adopted decay of autocorrelation.

In its present form the proposed framework is formulated for stationary processes. Therefore it can be directly used, for modeling of annual flows or short-timescale problems (e.g., rainfall generation within a rainfall event) that are not affected by seasonality. Thus it is not appropriate for problems involving periodic processes (e.g., seasonal flows). However, it can be directly linked to seasonal short-memory models such as the periodic autoregressive (PAR(1)) single-variate or multivariate model to simulate seasonal processes, as well. Such a linkage of annual to seasonal models has been studied elsewhere [*Koutsoyiannis and Manetas*, 1996]. The combination of an annual long-memory model and a seasonal model will preserve both the long-term memory properties, which will be indirectly transferred from the annual timescale into the seasonal timescale, and the seasonal properties. In addition, any other annual-to-seasonal disaggregation model [*Salas et al.*, 1980; *Ste-*

dinger and Vogel, 1984; *Grygier and Stedinger*, 1988, 1990; *Lane and Frevert*, 1990] could also be combined with the annual model in this respect.

Acknowledgments. The research leading to this paper was performed within the framework of the project "Modernization of the supervision and management of the water resource system of Athens," funded by the Water Supply and Sewage Corporation of Athens. The author wishes to thank the director, I. Nazlopoulos, the associate director, A. Xanthakis, and the members of the project committee for the support of the research. Thanks are also due to A. Efstratiadis for his comments and testing of the methodology. Finally, the author is grateful to the Associate Editor P. Kumar, the reviewers C. Puente and A. Carsteanu, and an anonymous reviewer for their constructive comments that were very helpful for an improved, more complete, and more accurate presentation.

References

Barnes, F. B., Storage required for a city water supply, *J. Inst. Eng. Aust.*, *26*(9), 198–203, 1954.

Beard, L. R., Use of interrelated records to simulate streamflow, *J. Hydraul. Div. Am. Soc. Civ. Eng.*, *91*(HY5), 13–22, 1965.

Bloomfield, P., *Fourier Analysis of Time Series*, John Wiley, New York, 1976.

Box, G. E. P., and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, Merrifield, Va., 1970.

Bras, R. L., and I. Rodriguez-Iturbe, *Random Functions in Hydrology*, Addison-Wesley-Longman, Reading, Mass., 1985.

Debnath, L., *Integral Transforms and Their Applications*, CRC Press, Boca Raton, Fla., 1995.

Ditlevsen, O. D., Extremes and first passage times, Doctoral dissertation, Tech. Univ. of Denmark, Lyngby, Denmark, 1971.

Eckhardt, R., Stan Ulam, John von Neumann and the Monte Carlo method, in *From Cardinals to Chaos*, edited by N. G. Cooper, Cambridge Univ. Press, New York, 1989.

Grygier, J. C., and J. R. Stedinger, Condensed disaggregation procedures and conservation corrections for stochastic hydrology, *Water Resour. Res.*, *24*(10), 1574–1584, 1988.

Grygier, J. C., and J. R. Stedinger, SPIGOT: A synthetic streamflow generation software package, version 2.5, technical description, Sch. of Civ. and Environ. Eng., Cornell Univ., Ithaca, New York, 1990.

Hurst, H. E., Long-term storage capacity of reservoirs, *Trans. Am. Soc. Civ. Eng.*, *116*, 776–808, 1951.

Koutsoyiannis, D., Optimal decomposition of covariance matrices for multivariate stochastic models in hydrology, *Water Resour. Res.*, *35*(4), 1219–1229, 1999a.

Koutsoyiannis, D., An advanced method for preserving skewness in single-variate, multivariate and disaggregation models in stochastic hydrology, paper presented at XXIV General Assembly of European Geophysical Society, The Hague, 1999b.

Koutsoyiannis, D., and E. Foufoula-Georgiou, A scaling model of storm hyetograph, *Water Resour. Res.*, *29*(7), 2345–2361, 1993.

Koutsoyiannis, D., and A. Manetas, Simple disaggregation by accurate adjusting procedures, *Water Resour. Res.*, *32*(7), 2105–2117, 1996.

Lane, W. L., and D. K. Frevert, Applied stochastic techniques, user's manual, personal computer version, Bur. of Reclam., Eng. and Res. Cent., Denver, Colo., 1990.

Maass, A., M. M. Hufschmidt, R. Dorfman, H. A. Thomas Jr., S. A. Marglin, and G. M. Fair, *Design of Water Resource Systems*, Harvard Univ. Press, Cambridge, Mass., 1962.

Mandelbrot, B. B., Une class de processus stochastiques homothetiques a soi: Application a la loi climatologique de H. E. Hurst, *C. R. Hebd. Seances Acad. Sci.*, *260*, 3284–3277, 1965.

Mandelbrot, B. B., A fast fractional Gaussian noise generator, *Water Resour. Res.*, *7*(3), 543–553, 1971.

Mandelbrot, B. B., and J. R. Wallis, Computer experiments with fractional Gaussian noises, 1, Averages and variances, *Water Resour. Res.*, *5*(1), 228–241, 1969a.

Mandelbrot, B. B., and J. R. Wallis, Computer experiments with fractional Gaussian noises, 2, Rescaled ranges and spectra, *Water Resour. Res.*, *5*(1), 242–259, 1969b.

Mandelbrot, B. B., and J. R. Wallis, Computer experiments with fractional Gaussian noises, 3, Mathematical appendix, *Water Resour. Res.*, *5*(1), 260–267, 1969c.

Marlow, W. H., *Mathematics for Operations Research*, Dover, Mineola, N. Y., 1993.

Matalas, N. C., Mathematical assessment of synthetic hydrology, *Water Resour. Res.*, *3*(4), 937–945, 1967.

Matalas, N. C., and J. R. Wallis, Generation of synthetic flow sequences, in *Systems Approach to Water Management*, edited by A. K. Biswas, McGraw-Hill, New York, 1976.

Mays, L. W., and Y.-K. Tung, Systems analysis, in *Water Resources Handbook*, edited by L. W. Mays, McGraw-Hill, New York, 1996.

McLeod, T. A., and K. W. Hipel, Simulation procedures for Box-Jenkins models, *Water Resour. Res.*, *14*(5), 969–975, 1978.

Mejia, J. M., I. Rodriguez-Iturbe, and D. R. Dawdy, Streamflow simulation, 2, The broken line process as a potential model for hydrologic simulation, *Water Resour. Res.*, *8*(4), 931–941, 1972.

Metropolis, N., The beginning of the Monte Carlo method, in *From Cardinals to Chaos*, edited by N. G. Cooper, Cambridge Univ. Press, New York, 1989.

O'Connell, P. E., Stochastic modelling of long-term persistence in streamflow sequences, Ph.D. thesis, Civ. Eng. Dep., Imperial Coll., London, 1974.

Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, 3rd ed., McGraw-Hill, New York, 1991.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, Cambridge Univ. Press, New York, 1992.

Salas, J. D., Analysis and modeling of hydrologic time series, in *Handbook of Hydrology*, edited by D. Maidment, chap. 19, pp. 19.1–19.72, McGraw-Hill, New York, 1993.

Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane, *Applied Modeling of Hydrologic Time Series*, Water Resour. Publ., Highlands Ranch, Colo., 1980.

Spiegel, M. R., *Theory and Problems of Laplace Transforms*, Shaum, New York, 1965.

Stedinger, J. R., and R. M. Vogel, Disaggregation procedures for generating serially correlated flow vectors, *Water Resour. Res.*, *20*(1), 47–56, 1984.

Thomas, H. A., and M. B. Fiering, Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation, in *Design of Water Resource Systems*, by A. Maass et al., Harvard Univ. Press, Cambridge, Mass., 1962.

Todini, E., The preservation of skewness in linear disaggregation schemes, *J. Hydrol.*, *47*, 199–214, 1980.

Wilson, G. J., Factorization of the generating function of a pure moving average process, *SIAM J. Numer. Anal.*, *6*, 1, 1969.

D. Koutsoyiannis, Department of Water Resources, Faculty of Civil Engineering, National Technical University, Athens, Heroon Polytechneiou 5, GR-157 80 Zographou, Greece.

(Received July 12, 1999; revised February 15, 2000;
accepted February 18, 2000.)